文章编号：2095-2163(2022)07-0008-07 中图分类号：R541.4;TP181 文献标志码：A

基于集成学习的冠心病风险预测模型研究

苏文星[1]，张振一[1]，郑琰莉[3]，唐 琳[4]，宋元涛[2]

（1 中国科学院大学 工程与科学学院，北京 100049；2 中国科学院大学 应急管理科学与工程学院，北京 100049；
3 天津泰达普华医院，天津 300457；4 西安理工大学 经济与管理学院，西安 710045）

摘 要：近年来，冠心病患者人数不断增加，而集成学习具有良好的冠心病风险预测能力，可降低患者就医成本，提高冠心病筛查的效率。本文利用 Kaggle 平台公开的冠心病数据集，首先对数据集进行了预处理和特征指标筛选，并利用 SMOTE 算法对数据进行类别平衡，最终得到 7 010 条数据；选取随机森林、XGBoost、LightGBM 3 个集成学习算法，构建相应的冠心病风险预测模型，并利用贝叶斯优化算法对模型进行超参数调优，同时将数据以 7∶3 的比例分为训练集与测试集进行模型训练与预测；最后，通过准确率、召回率、*AUC* 等指标对 3 种模型的性能进行比较。结果显示 3 种集成学习算法预测模型性能均较好，其中 LightGBM 算法预测模型性能最为突出，验证了集成学习算法运用在冠心病风险预测方面的可行性。

关键词：冠心病；集成学习；贝叶斯优化；SMOTE；风险预测模型

Research on risk prediction model of coronary heart disease based on ensemble learning

SU Wenxing[1], ZHANG Zhenyi[1], ZHENG Yanli[3], TANG Lin[4], SONG Yuantao[2]

(1 School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China;
2 School of Emergency Management Science and Engineering, University of Chinese Academy of Sciences, Beijing 100049, China;
3 Tianjin TEDA Puhua International Hospital, Tianjin 300457, China;
4 College of Economics and Management, Xi′an University of Technology, Xi′an 710045, China)

【Abstract】 In recent years, the number of people suffering from coronary heart disease is increasing. Ensemble learning has excellent prediction ability for coronary heart disease, which can reduce the cost of patient care and improve the efficiency of coronary heart disease screening. The datasets for this study have been published by Kaggle. Firstly, the data is preprocessed and screened with characteristic indexes, and the SMOTE algorithm is used to balance the data categories after which eventually 7010 pieces of data are obtained. Secondly, three ensemble learning algorithms, Random Forest, XGBoost, and LightGBM are selected to construct the corresponding coronary heart disease risk prediction model, and the Bayesian optimization algorithm is used to optimize the hyperparameters of the model. At the same time, the data is divided into training set and test set in a ratio of 7/3 for model training and prediction. Finally, the performance of the three models is compared by accuracy, recall, *AUC* and other metrics. The results show that the prediction models of the three ensemble learning algorithms all have good performance, among which the LightGBM algorithm has the most prominent performance, which verifies the feasibility of the ensemble learning algorithm in the risk prediction of coronary heart disease.

【Key words】 coronary heart Disease; ensemble learning; Bayesian optimization; SMOTE; risk prediction model

0 引 言

冠状动脉粥样硬化性心脏病简称“冠心病”(Coronary Heart Disease, CHD)，是指冠状动脉血管发生动脉粥样硬化病变而引起血管腔狭窄或阻塞，造成心肌缺血、缺氧或坏死而导致的心脏疾病。随着老龄化进程加快以及居民不良生活方式的影响，心血管疾病的发病率逐年增高。中国患有心血管病的人数约为 3.3 亿，其中冠心病 1 139 万人，且农村地区心血管病死亡率持续高于城市水平[1]。目前，临床上对冠心病诊断主要依靠临床症状、实验室检查、影像学检查诊断等，其中冠状动脉造影(Coronary Angiography, CAG)是诊断冠心病的“金标准”，但诊断过程繁琐且费用较为昂贵[2]。如能早期对冠心病给予相应的风险预测，可在降低居民患病风险和就医成本的同时提高疾病筛查的效率，因

作者简介：苏文星(1995-)，男，硕士研究生，主要研究方向：信息系统工程管理；张振一(1990-)，男，硕士研究生，主要研究方向：工程管理；郑琰莉(1995-)，女，大专，初级护师，主要研究方向：心内科护理学；唐 琳(1999-)，女，硕士研究生，主要研究方向：供应链管理；宋元涛(1974-)，男，博士，副教授，主要研究方向：人工智能、应急管理、工程管理。

通讯作者：宋元涛 Email：songyuantao@ucas.ac.cn

收稿日期：2022-01-12

此，找到快速又经济的冠心病早期预测方法具有重要意义。

近年来，机器学习由于其强大的数据分类与预测能力，在疾病预测及辅助临床治疗决策方面做出一定贡献，集成学习算法的预测效果尤为突出[3]，但机器学习在冠心病风险预测方面并未得到广泛应用。此外，相关研究发现，高血压、高胆固醇、糖尿病以及年龄、性别、身体质量指数（BMI）、是否吸烟等都会影响患冠心病的几率[4]。因此，本文利用 Kaggle 平台公开的 CHD 数据集，基于随机森林、XGBoost、LightGBM 3 种较为成熟的集成学习（Ensemble learning）算法建立冠心病风险预测模型，利用准确率、召回率、*AUC* 等指标对 3 种模型的性能进行比较，验证集成学习算法在冠心病风险预测方面的可行性，从而实现对冠心病的早期风险预测。

1　数据处理与特征工程

1.1　数据来源

本文数据源为 Kaggle 官方大数据平台提供的针对马萨诸塞州弗雷明翰镇居民心血管研究公开数据，其分类目标是预测患者 10 年间是否罹患冠心病，如果有计作 1（阳性），否则计作 0（阴性）。数据集共有 4 283 条记录，涵盖了人口统计学、行为学和医学风险 3 个维度的 15 个风险特征指标。具体特征指标变量见表 1。

表 1　风险特征指标变量详情与解释

Tab. 1　Detail and explanation for risk characteristic index variables

指标维度	特征指标名称	变量名称	赋值解释
人口统计学	性别	Male	0 =女性；1 =男性
	年龄	Age	实际年龄（数字）
行为学	教育程度	Education	1 =高中以下；2 =高中；3 =职业学校；4 =大学
	是否吸烟	currentSmoker	0 =不吸烟者；1 =吸烟者
	平均每日吸烟量	cigsPerDay	实际平均每日吸烟数量（数字）
医学风险（历史）	是否服用降压药	BPMeds	0 =不服用降压药；1 =服用降压药
	中风史	prevalentStroke	0 =否；1 =是
	高血压史	prevalentHyp	0 =否；1 =是
	糖尿病史	Diabetes	0 =否；1 =是
医学风险（当前）	总胆固醇水平	totChol	实际测量数值
	收缩压	sysBP	实际测量数值
	舒张压	diaBP	实际测量数值
	身体质量指数	BMI	实际测量数值
	心率	heartRate	实际测量数值
	血糖水平	Glucose	实际测量数值

1.2　数据分析与缺失值处理

采用 Pandas 对数据源数据对指标变量的值类型、分布以及缺失情况进行分析得出：数据不满足正态分布（$p < 0.05$），且教育程度（education）、平均每日吸烟量（cigsPerDay）、是否服用降压药（BPMeds）、总胆固醇水平（totChol）、身体质量指数（BMI）、血糖水平（glucose）存在数据的缺失。具体特征指标变量数据情况见表 2。

表 2　特征指标变量数据情况

Tab. 2　Data of characteristic index variables

变量名称	变量解释	非空值数	空值数	空值占比/%	值类型
male	性别	4 238	0	0.00	int64
Age	年龄	4 238	0	0.00	int64
education	教育程度	4 133	105	2.48	float64
currentSmoker	是否吸烟	4 238	0	0.00	int64
cigsPerDay	平均每日吸烟量	4 209	29	0.68	float64
BPMeds	是否服用降压药	4 185	53	1.25	float64
prevalentStroke	中风史	4 238	0	0.00	int64
prevalentHyp	高血压史	4 238	0	0.00	int64
diabetes	糖尿病史	4 238	0	0.00	int64
totChol	总胆固醇水平	4 188	50	1.18	float64
sysBP	收缩压	4 238	0	0.00	float64
diaBP	舒张压	4 238	0	0.00	float64
BMI	身体质量指数	4 219	19	0.45	float64
heartRate	心率	4 237	1	0.02	float64
glucose	血糖水平	3 850	388	9.16	float64

数据的缺失会影响数据分析的质量和建模的准确性,所以需要针对不同特征变量数据分析情况采取恰当方式进行数据处理。教育程度指标变量受患者实际情况影响,数据不可得且缺失数据的比例在5%以下,可以使用删除法对缺失值进行处理[5];平均每日吸烟量的缺失值,分析发现对应记录均为吸烟者,因此取所有吸烟者且每日吸烟量非空数据的平均数(18.0)对缺失值进行插值;对于是否服用降压药指标变量缺失值,参考美国心脏病协会(American Heart Association,AHA)高血压指南最新诊断标准,在未使用降压药物的情况下,收缩压(systolic blood pressure,SBP)≥130 mmHg 和(或)舒张压(Diastolic Blood Pressure,DBP)≥80 mmHg 的人群诊断为高血压患者,对收缩压大于 130 mmHg 或者舒张压大于 80 mmHg 的数据以 1 进行插值,否则以 0 进行插值[6];对于总胆固醇水平、身体质量指数、心率和血糖水平指标变量的缺失数据,其数据比例均占总数据 10%以下,分别求各指标变量数据平均值后对空缺数据进行填补[7]。本文基于 Python 的 pandas 工具库对上述数据进行处理,最终得到4 133条样本用于模型构建,其中阴性患者 3 505例(84.8%),阳性患者 628 例(15.2%)。部分样本数据见表 3。

表 3 部分样本数据

Tab. 3 Part of sample data

指标	指标说明	序号				
		01	02	58	59	99
Male	性别	1	0	0	1	0
Age	年龄	39	46	55	48	39
education	教育程度	4	2	1	1	2
currentSmoker	是否吸烟	0	0	0	1	1
cigsPerDay	平均每日吸烟量	0	0	0	15	20
BPMeds	是否服用降压药	0	0	0	0	0
prevalentStroke	中风史	0	0	0	0	0
prevalentHyp	高血压史	0	0	0	0	0
diabetes	糖尿病史	0	0	0	0	0
totChol	总胆固醇水平	195	250	330	170	190
sysBP	收缩压	106	121	103	132	137
diaBP	舒张压	70	81	73	91	81
BMI	身体质量指数	26.97	28.73	24.5	27.61	19.57
heartRate	心率	80	95	85	78	80
glucose	血糖水平	77	76	67	57	85
TenYearCHD	是否罹患冠心病	0	0	0	1	1

1.3 特征分析与选择

特征选择旨在通过分析特征间的关系筛选出对模型贡献度较高的特征变量,以提高模型的性能。鉴于数据不满足正态分布,本文首先基于 Spearman 秩相关系数对特征指标变量相关性进行分析,具体相关情况如图 1 所示。其中年龄(age)、收缩压(sysBP)、是否患有高血压(prevalentHyp)、舒张压(diaBP)、血糖水平(glucose)为重要特征,是否吸烟-平均每日吸烟(currentsmoker-cigsperday)的相关系数为 0.93,舒张压-收缩压(diaBP-sysBP)的相关系数为 0.78,高血压史-收缩压(prevalentHyp-sysbp)的相关系数为 0.70,高血压史-舒张压(prevalentHyp-diaBP)的相关系数为 0.62,P 值均小于 0.05,特征指标变量间存在较高相关性。分析可得特征指标变量与目标值相关性均小于 0.6 且特征指标数量较少,故保留所有特征指标变量进行模型预测。

一般来说,不平衡数据集会削弱学习算法预测准确性,本文应用的冠心病数据集中阳性与阴性数据比值约为 1∶6,数据类别不平衡明显。人工少数类过采样法(Synthetic Minority Over-Sampling Technique,SMOTE)在解决数据类别不平衡问题上具有良好的效果[8]。本文将采用该方法随机生成新实例以平衡数据。

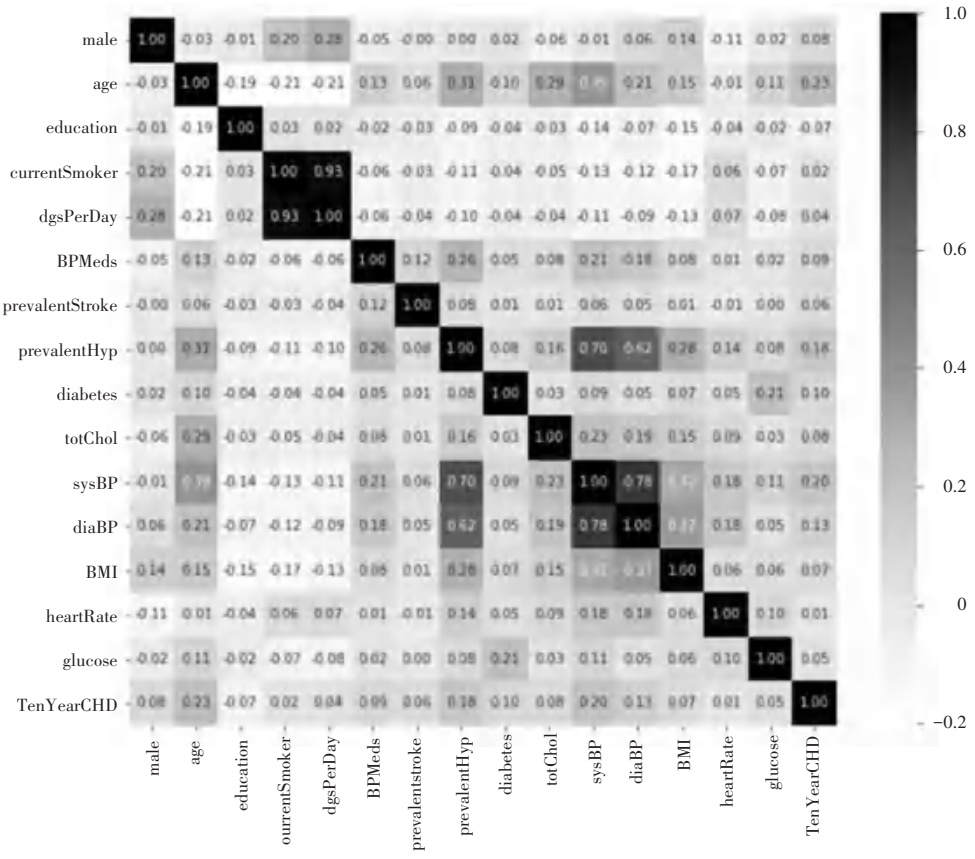

图 1　特征指标变量相关性

Fig. 1　Correlation between characteristic index variables

2　风险预测模型构建

2.1　模型介绍

集成学习通过构建并结合多个学习器来完成学习任务，通过多分类器的预测结果来改善基本学习器的泛化能力和鲁棒性。本文选取随机森林、XGBoost、LightGBM 3 种较为成熟的集成学习算法建立冠心病风险预测模型。

（1）随机森林：随机森林具有易于实现、抗噪声能力优、数据集适应能力强且不易陷入过拟合等特点。随机森林（Random Forest，RF）利用集成学习的思想，包含多个决策树的分类器，对于一个输入样本，不同树会有不同的分类结果，随机森林通过随机的方式建立多棵决策树并集成了所有的分类投票结果，选择投票数最多的类别作为最终的输出。

（2）XGBoost：XGBoost（eXtreme Gradient Boosting）是一种改进的梯度提升算法，具有计算复杂度低、灵活性强、运行速度快，精准度高的优点。基于预排序方法并通过对误差函数进行二阶泰勒展开，加入正则化项来优化目标函数，将多个弱分类器进行融合，从而形成强分类器，同时采用收缩（Shinkage）、列特征抽样（Column Subsampling）等方法来防止过拟合。

（3）LightGBM：LightGBM（Light Gradient Boosting Machine）是基于决策树算法的分布式梯度提升框架，采用 Histogram 算法，使用带有深度限制的按叶子生长（leaf-wise）策略，支持高效率的并行训练，并且具有更快的训练速度、更低的内存消耗、更好的准确率。支持分布式可以快速处理海量数据，训练效果好、不易过拟合。3 种集成学习算法在模型设计上存在明显差异，其核心的不同点见表 4。

2.2　模型参数选择

为了确保模型有较好的效果，需要对模型调参，手动调参十分耗时且依赖于个人经验，网格和随机搜索调参需要很长的运行时间。贝叶斯优化法目前广泛应用于解决机器学习中的超参数搜索问题，同时该方法较随机搜索具有省时、性能优的特点[9]。因此，本文基于 Python 语言使用 BayesianOptimization 库

对 3 种模型使用贝叶斯优化法进行超参数调优。其中 n_estimator 代表建立子树的数量,一般来说模型的性能与子树的数量成正比,但是数值过大可能会导致模型过拟合,因此随机森林、XGboost、LightGBM3 种集成学习模型基于贝叶斯优化的优化结果分别为 383/398/574,其他参数的详细设置情况见表 5~表 7。

2.3 模型预测结果

确定好模型参数后,本文基于 Python 语言并结合 sklearn 机器学习库,首先将数据集按照 7∶3 的比例分割为训练数据集和测试数据集,在训练数据集上完成 3 个模型的训练,并使用训练好的模型在测试数据集上测试,得到相应的预测结果(混淆矩阵),见表 8。

表 4 3 种集成学习算法对比表

Tab. 4 Comparison of three ensemble learning algorithms

算法	模型思想	切分算法	决策树生长策略	增益计算
随机森林	Bagging	随机切分	Gini 系数	Gini 系数
XGBoost	Boosting	pre-sorted	level-wise	优化推导公式
LightGBM	Boosting	histogram	leaf-wise	优化推导公式

表 5 随机森林算法模型参数设置情况

Tab. 5 Parameter settings of random forest model

参数	n_estimators	max_depth	max_features	min_samples_split	min_samples_leaf
释义	子树的数量	树的最大深度	最大特征数	分裂所需的最小样本数	叶节点最小样本数
值	383	20	4	2	1

表 6 XGboost 算法模型参数设置情况

Tab. 6 Parameter settings of XGboost model

参数	booster	Objective	n_estimators	max_depth	learning_rate	colsample_bytree	subsample
释义	分类器类型	目标函数	子树的数量	树的最大深度	学习率	特征采样比例	子样本占比
值	bgtree	binary: logistic	398	14	0.2	0.8	0.9

表 7 LightGBM 算法模型参数设置情况

Tab. 7 Parameter settings of LightGBM model

参数	boosting_type	objective	n_estimators	max_depth	learning_rate	colsample_bytree	subsample	num_leaves
释义	分类器类型	目标函数	子树的数量	树的最大深度	学习率	特征采样比例	子样本占比	叶子节点数
值	gbdt	binary	574	20	0.2	0.9	0.9	196

表 8 3 种模型的预测结果(混淆矩阵)

Tab. 8 Prediction results of three models (confusion matrix)

结果	LightGBM		XGboost		随机森林	
	阳性(预测)	阴性(预测)	阳性(预测)	阴性(预测)	阳性(预测)	阴性(预测)
阳性	945	123	932	136	951	117
阴性	61	974	77	958	86	949

3 模型性能评价与比较

3.1 模型性能评价指标

本文主要以准确率(*Accuracy*,*ACC*)、精确率(*Precision*,*PRE*)、召回率(*Recall*)、F_1 值和 *AUC* 值评估算法的适用性及效果,同时使用 10 折交叉验证(*K*10)的方式验证模型的性能。在解释上述评价指标之前需要对混淆矩阵进行释义,首先把预测值与实际值两两匹配,然后显示预测结果为阳性/阴性(Positive/Negative),再根据实际与预测结果对比,得出判断结果为正确/错误(True/False),最终得到混淆矩阵见表 9。

表 9 混淆矩阵

Tab. 9 Confusion matrix

实际值	预测值	
	阳性	阴性
阳性	*TP* (True Positive)	*FN* (False Negative)
阴性	*FP* (False Positive)	*TN* (True Negative)

准确率是指预测模型预测正确的结果占总样本的百分比，计算公式(1)：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

精准率是指在所有被预测为阳性的样本中实际值也为阳性的样本所占百分比，计算公式(2)：

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

召回率是指在所有实际值为阳性的样本中被预测为阳性的样本所占百分比，计算公式(3)：

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F_1 值是为了更好的进行整体评价，在 *Precision* 和 *Recall* 的基础上，使用两者的加权调和平均进行模型性能效果的评价，计算公式(4)：

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

除此之外，本文还引入 *ROC*（Receiver Operating Characteristic Curve）曲线对模型进行评估，*ROC* 是以假阳性率（False Positive Rate）、真阳性率（True Positive Rate）为轴的曲线，*ROC* 曲线下的面积（Area Under Curve，*AUC*）可以直观的评价分类器的好坏，范围在 0~1 之间，值越大代表模型性能越好。

3.2　模型比较结果

利用 Python 语言 metrics 库得出 3 种预测模型的性能度量结果，见表 10，可以看出：3 种算法的准确率均在 90%左右且数值相差不大，预测效果均较为良好；相较于其他模型，LightGBM 的精准度最高，为 93.94%；由表 10 和图 2 可以看出，3 种算法的 *AUC* 值均在 0.9 以上且 3 种算法 10 折交叉验证的准确率均在 85%左右，表明其准确性、稳定性均较好。从 F_1 值指标上观察，LightGBM 模型预测效果略优于其他 2 个模型。综合上述指标可以看出，在本次选取的 3 种模型的训练效果均较好，LightGBM 性能最为优秀。本文通过与相关研究成果对比发现，本研究选取的 3 种模型在准确率与 *AUC* 值方面较其有明显提升。

表 10　3 种模型性能度量指标对比

Tab. 10　Comparison of performance metrics of three models

算法	*ACC*/%	*PRE*/%	*Recall*/%	F_1 值	*AUC*	K10/%
随机森林	90.35	91.71	89.04	90.36	0.904	83.55
XGBoost	89.87	92.37	87.27	89.74	0.899	84.64
LightGBM	91.25	93.94	88.48	91.13	0.913	85.45

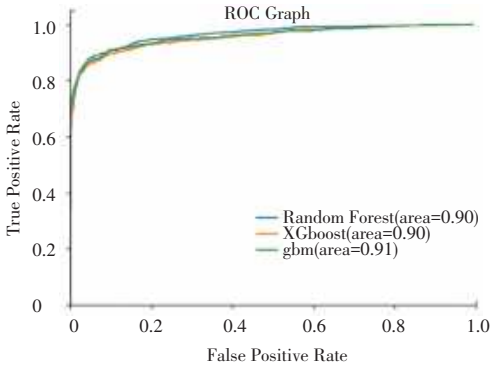

图 2　3 种模型的 *ROC* 曲线

Fig. 2　*ROC* curves of three models

4　结束语

冠心病是最常见的心血管疾病之一，而现阶段的诊疗成本较高，如能早期对冠心病给予相应的风险预测，提高疾病筛查的效率，不仅可降低居民的患病风险，还可降低患者就医成本，因此选择科学有效的方法进行早期冠心病的风险预测是非常有意义的。本文基于 Kaggle 上公开的冠心病数据集，首先对数据进行分析并对缺失数据按照不同情况处理，并利用 SMOTE 算法对数据进行平衡处理；采用随机森林、XGboost、LightGBM 3 种集成学习算法模型构建了冠心病的风险预测模型，并使用贝叶斯优化算法对模型进行了调优；最后，从准确率、召回率、*AUC* 等指标对 3 种模型的性能进行比较，发现 3 种模型均具有良好的性能，验证集成学习算法在冠心病风险预测方面的可行性，从而实现冠心病早期风险预测。此外，基于机器学习建立的风险预测模型不仅可以对冠心病进行风险预测，还可以将其推广到预测其他类型的疾病，以提高疾病的早期筛查效率。

本文也存在一定的局限性。首先，本文采用的数据来源于开放平台，在数据数量、质量以及适用性上存在一定的局限性，未来考虑使用医院的真实大数据进行模型构建与预测；其次，本文使用的算法模型均为集成学习范畴，以后可考虑选取不同类型的机器学习算法进行改进对比，构建更加优秀的风险预测模型。

参考文献

[1] 中国心血管健康与疾病报告 2020[J]. 心肺血管病杂志，2021，40(9)：885-889.

[2] 胡大一. 冠心病诊断与治疗研究进展[J]. 中华心血管病杂志，2003(11)：9-14.

（下转第 19 页）